Quantitative EEG reactivity and machine learning for prognostication in hypoxic-ischemic brain injury ☆

Edilberto Amorim [a,b,*], Michelle van der Stoel [c], Sunil B. Nagaraj [d], Mohammad M. Ghassemi [e], Jin Jing [a], Una-May O'Reilly [b], Benjamin M. Scirica [f], Jong Woo Lee [g], Sydney S. Cash [a], M. Brandon Westover [a]

[a] Department of Neurology, Massachusetts General Hospital, Boston, MA, USA
[b] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA
[c] University of Twente, Enschede, Netherlands
[d] University of Groningen, Groningen, Netherlands
[e] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA
[f] Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA
[g] Department of Neurology, Brigham and Women's Hospital, Boston, MA, USA

ARTICLE INFO

HIGHLIGHTS

- Machine-learning methods using QEEG reactivity data can predict outcomes after cardiac arrest.
- A QEEG reactivity detector can provide individual-level predictions of neurological recovery.
- A quantitative approach to prognostication may improve objectivity of EEG reactivity interpretation.

ABSTRACT

Objective: Electroencephalogram (EEG) reactivity is a robust predictor of neurological recovery after cardiac arrest, however interrater-agreement among electroencephalographers is limited. We sought to evaluate the performance of machine learning methods using EEG reactivity data to predict good long-term outcomes in hypoxic-ischemic brain injury.
Methods: We retrospectively reviewed clinical and EEG data of comatose cardiac arrest subjects. Electroencephalogram reactivity was tested within 72 h from cardiac arrest using sound and pain stimuli. A Quantitative EEG (QEEG) reactivity method evaluated changes in QEEG features (EEG spectra, entropy, and frequency features) during the 10 s before and after each stimulation. Good outcome was defined as Cerebral Performance Category of 1–2 at six months. Performance of a random forest classifier was compared against a penalized general linear model (GLM) and expert electroencephalographer review.
Results: Fifty subjects were included and sixteen (32%) had good outcome. Both QEEG reactivity methods had comparable performance to expert EEG reactivity assessment for good outcome prediction (mean AUC 0.8 for random forest vs. 0.69 for GLM vs. 0.69 for expert review, respectively; *p* non-significant).
Conclusions: Machine-learning models utilizing quantitative EEG reactivity data can predict long-term outcome after cardiac arrest.
Significance: A quantitative approach to EEG reactivity assessment may support prognostication in cardiac arrest.

1. Introduction

Most patients resuscitated from a cardiac arrest who undergo targeted temperature management are initially comatose (Nielsen et al., 2013). Lack of early clinical exam improvement following resuscitation may influence providers' perceptions of neurological prognosis and lead to inappropriately premature

withdrawal of life-sustaining therapies (Perman et al., 2012). Identification of early predictors of neurological recovery has the potential to promote continuation of intensive life support measures and prevent self-fulfilling prophecies due to premature withdrawal of life-sustaining therapies in cardiac arrest patients with unfavorable prognostic features identified on multimodal prognostication but who may have the potential for good outcomes if enough time for neurological recovery is provided.

Electroencephalogram (EEG) reactivity is defined as a "change in cerebral activity to external stimulation" (Hirsch et al., 2013). This phenomenon has been well-studied in the clinical neurophysiology and coma literature, but only recently has been systematically investigated in large cohorts of patients with acute hypoxic-ischemic brain injury after cardiac arrest (Amorim et al., 2016a, 2015, 2016b; Rossetti et al., 2010, 2017; Westhall et al., 2016) In these studies, EEG reactivity was one of the most compelling predictors of long-term neurological outcome, being often observed within the first 24 h from return of spontaneous circulation despite the presence of hypothermia, sedation, and a clinical exam consistent with coma (Amorim et al., 2016a; Rossetti et al., 2017).

Despite strong evidence supporting the use of EEG reactivity in cardiac arrest prognostication, its clinical usability is limited by poor expert interrater agreement and high variability in visual assessment of EEG reactivity practices (Amorim et al., 2018; Fantaneanu et al., 2016; Hermans et al., 2016; Westhall et al., 2016) Advancements in computational analysis of EEG data have shown that quantitative approaches can begin to overcome these challenges and set the stage for bedside technologies that can deliver objective and real-time quantification of EEG reactivity that is comparable or even superior to human expert performance (Bricolo et al., 1978; Duez et al., 2018; Hermans et al., 2016; Liu et al., 2016; Noirhomme et al., 2014).

In this study, our primary aim was evaluate whether a machine-learning algorithm designed to assess QEEG reactivity could predict good long-term outcomes in comatose cardiac arrest subjects undergoing targeted temperature management. We devised a method with flexibility to allow stimulus specific (sound or pain) and stimulus agnostic (sound and pain) assessments of QEEG reactivity, and also with the added capability of providing individualized long-term outcome predictions for each subject tested using a QEEG reactivity-based risk score. We compare our machine learning QEEG reactivity method performance predicting long-term outcomes to expert visual assessment of EEG reactivity.

2. Methods

2.1. Subjects, EEG acquisition, and medical management

Adult subjects who had return of spontaneous circulation (determined by a healthcare provider and recorded in the electronic medical records) and remained comatose following an in-hospital or out-of-hospital cardiac arrest from December 2011 to April 2016 in two university-affiliated hospitals were screened. Continuous digital EEG monitoring according to the international 10–20 system is routinely started as early as possible during targeted temperature management and continued for 24–72 h. Monitoring with EEG may be discontinued prior to 72 h in case of return of consciousness, withdrawal of life-sustaining therapies, or death.

At the time of this study, the targeted temperature management protocol at participating hospitals utilized external cooling pads and had a goal temperature of 32–34 °C for 24 h followed by slow rewarming at 0.25–0.5 °C per hour to a goal temperature of 37 °C. Neuromuscular blockade was maintained during the entirety of the hypothermia phase in one of the hospitals and utilized as needed for shivering management in the other. Propofol, midazolam, or fentanyl infusions were utilized for sedation and analgesia per local institution protocol and were titrated according to the treating clinicians' discretion.

Data collection and analysis was approved by the Partners Healthcare Institutional Review Board, which deemed this retrospective analysis of demographic, clinical, and EEG data exempt from requirements for informed consent.

2.2. EEG reactivity testing battery and expert visual review

As part of standard clinical care and per local institutional protocol, EEG reactivity is tested by the clinical neurology team during hypothermia and normothermia when possible. The EEG reactivity testing procedure included sound, tactile, or pain stimulation. All subjects had EEG reactivity scored as "present", "absent", or "indeterminate" by a board-certified expert electroencephalographer during clinical care. The EEG was scored as having reactivity present if a change in EEG signal amplitude or frequency was observed. An epoch with duration of 10 s was used as most EEG reactivity responses evaluated by electroencephalographers are observed within this time frame in our experience. If EEG reactivity was present in any of the first three days of recording, the expert final score used in our analysis was "EEG reactivity present". Activation procedures such as calling the subject's name or clapping were categorized as "sound stimuli"; tactile stimulation at the torso, trapezius pressure, nail bed pressure, sternal rub, nose stimulation with a swab, or tracheal suction were categorized as "pain stimuli".

2.3. Data collection and functional outcome assessment

Electronic medical records were reviewed retrospectively and information regarding age, gender, time to return of spontaneous circulation (ROSC), initial cardiac rhythm – dichotomized as shockable (ventricular fibrillation or ventricular tachycardia) or non-shockable (asystole, pulseless electrical activity, and unknown) – admission Glasgow Coma Score, pupillary and corneal exam, admission computerized tomography of the head (CTH) reports, presence of malignant EEG patterns within the first 72 h from ROSC (burst-suppression, seizure, status epilepticus, myoclonic status epilepticus, or periodic discharges), reason for withdrawal of life-sustaining therapies (WLST), and best neurological exam prior to discharge was retrieved. A CTH radiology report describing acute loss of cortical or subcortical gray-white matter differentiation was categorized as unfavorable. Subjects in which WLST was motivated by predicted poor neurologic outcome were categorized as "WLST due to Poor Neurological Prognosis". Other causes of WLST due to "non-neurological" reasons such as multi-organ failure or refractory hemodynamic instability, or other causes of death such as brain death were not categorized as WLST due to Poor Neurological Prognosis. Arousal recovery was defined as at least recovery to minimal conscious state with command following or visual tracking (Giacino et al., 2002).

Functional outcome was defined as the best neurological function achieved up to 6-months after initial cardiac arrest utilizing the Glasgow-Pittsburgh Cerebral Performance Categories Scale (CPC) (Safar, 1981). "Good" functional outcome was defined as a CPC score of 1 or 2 and "poor" functional outcome was defined as a CPC of 3–5. The CPC score was determined based on electronic medical record review of patients who had not already achieved a good functional outcome or died by the time of discharge, i.e. subjects discharged with CPC 3 or 4.

2.4. EEG data pre-processing

The EEG was re-referenced to a bipolar montage and baseline drift was removed. All EEG data was filtered (0.5–50 Hz) and channels with the following types of artifact: were excluded: amplitude >500 µV or low variance using a custom made algorithm. (Fig. 1A, preprocessing) Twenty-second long EEG clips were obtained for each clinical stimulation documented done during EEG reactivity testing. Each 20-s long EEG clip was separated in two 10-s long epochs centered on the time of external stimulation (Fig. 1A, pre and post-stimulation epochs). Epochs used for EEG reactivity assessment that had burst suppression, epileptiform discharges, or prominent artifacts were excluded from analysis. This visual screening was performed by a fellowship-trained electroencephalographer (E.A.) using MATLAB v2017a (Natick, MA).

2.5. Quantitative EEG feature extraction and post-processing

Sixty-two quantitative EEG (QEEG) features assessing frequency, signal complexity, and entropy were utilized to evaluate the EEG clips (Table 1). Quantitative EEG features were extracted for each epoch and each channel independently. The QEEG features were averaged all channels for each feature and across all 10-s pre and post-stimulus epochs separately. The power spectrum of the EEG clips was estimated with Thomson's multi-taper method using the Chronux toolbox (Bokil et al., 2010; Thomson, 1982). Subsequently, for each individual patient, EEG clips were separated in two categories based on the type of activation procedure performed (sound or pain stimuli). Subsequently, all QEEG data for each of the activation procedure categories were averaged (for each feature separately), creating a unique mean QEEG feature value for the before and after stimulation epochs for each feature (Fig. 1A, average QEEG Features per Stimulus Category). The absolute difference between pre and post-stimulation mean QEEG feature values was used as the measure of QEEG reactivity response for each feature for each individual patient. The QEEG features were normalized for each cross-validation fold in the training set, i.e. subtracting the mean QEEG feature value across patients and dividing it by the standard deviation. The QEEG features in the testing set were normalized based on the mean and standard deviation of the training set. The same procedure was done for the stimulus agnostic, but QEEG data from both sound and pain stimuli were

Table 1
Summary of quantitative EEG features acquired for EEG reactivity assessment.

Feature domain	Quantitative EEG features
Complexity/ Entropy	Nonlinear Energy Operator (Mukhopadhyay and Ray, 1998); Hjorth Parameter (Activity; Mobility; Complexity) (Hjorth, 1970); Fractal Dimension (Lutzenberger et al., 1992); Singular Value Decomposition Entropy (Sabatini, 2000); Spectral Entropy (Zhang et al., 2001); State/Response Entropy (Lysakowski et al., 2009); Sample Entropy (Abasolo et al., 2006); Renyi Entropy (Kannathal et al., 2005); Shannon Entropy (Kannathal et al., 2005); Approximate Entropy (Liang et al., 2015); Permutation Entropy (Bandt and Pompe, 2002); Relative Entropy (Inouye et al., 1991); Kurtosis; Skewness
Amplitude	Root Mean Square Amplitude; Mean of Amplitude Modulation; Standard Deviation of Amplitude Modulation; Skewness of Amplitude Modulation; Kurtosis of Amplitude Modulation; Mean of Amplitude Modulation (Stevenson et al., 2013); Burst Suppression Ratio (Nagaraj et al., 2018)
Frequency	Band Power: Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–12 Hz), Spindle (12–16 Hz), and Beta (16–25 Hz) Band Power; Total Power (0.5–32 Hz) Band Power Normalized by Total Power: Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–12 Hz), Beta (16–25 Hz) Band Power Ratios: Theta/Delta, Alpha/Delta, Spindle/Delta, Beta/Delta, Alpha/Theta, Spindle/Theta, and Beta/Theta Brain Symmetry Index (BSI) (van Putten et al., 2004): BSI total; BSI alpha; BSI delta; BSI theta; BSI beta (12–17 Hz); BSI beta (18 Hz); BSI delta and alpha combined; BSI delta, theta, and alpha combined Two Group Test (TGT) (Bokil et al., 2007): TGT total; TGT alpha; TGT delta; TGT theta; TGT beta (12–17 Hz); TGT beta (18 Hz); TGT delta and alpha combined; TGT delta, theta, and alpha combined Standard Deviation of Frequency Modulation; Skewness of Frequency Modulation; Kurtosis of Frequency Modulation (Stevenson et al., 2013); Kolmogorov-Smirnov (spectral distribution) (McEwen and Anderson, 1975); Spectral Edge Frequency; Peak Frequency

References are listed in the Supplementary Material 1.

averaged together. For univariate analyses, we averaged the QEEG feature values from all activation procedure categories, i.e. sound and pain. In case subjects had EEG reactivity assessment in consecutive days, QEEG features were only obtained from the first EEG reactivity assessment. Data from additional days of EEG reactivity
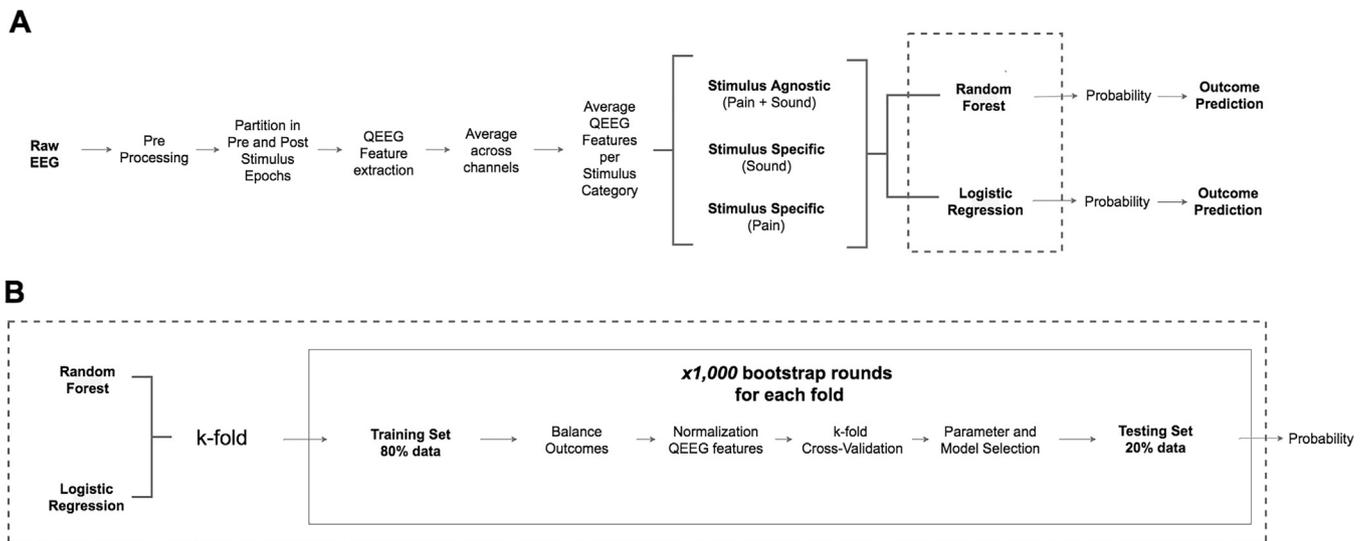

Fig. 1. A: Architecture data processing and evaluation: preprocessing, feature acquisition, separation by stimulus type; B: Classification steps for random forest model and penalized logistic regression.

assessment were discarded. This procedure aims to generate a more uniform outcome prediction procedure, as including data from several days could bias EEG reactivity assessment as individual responses may change over time.

2.6. Statistical analysis

Univariate analysis for good outcome prediction was performed using Spearman's rank coefficient for all individual QEEG reactivity features, Pearson X^2 for categorical variables, and independent t-Student tests for continuous variables. We evaluated the multivariate model performance for good outcome prediction using the area under the receiver operator curve (AUC).

Two multivariate models were utilized to evaluate good outcome prediction using all QEEG reactivity features (random forest classifier and penalized multinomial logistic regression with parameter optimization). The random forest classifier is a method known to provide highly accurate classifications, and it is especially useful when features may be correlated. This method also provides estimates of what variables are more relevant in the classification. Binary logistic regression is a commonly used classifier method that provide probability scores for observations, in this study, the odds of good outcome given a set of predictors. A penalized logistic regression with parameter optimization improves prediction accuracy and interpretability and also avoids overfitting by performing feature selection and regularization. It penalizes features with large regression coefficients, leading to a simpler model composed of the subset of features with best prediction performance. The EEG data was divided based on the activation procedure category (sound and pain) for the **stimulus specific** analysis, and data from each activation procedure group was analyzed separately. The **stimulus agnostic** analysis combined the EEG data recorded during both sound and pain activation procedures. Clinical information included in the prediction model in addition to the QEEG reactivity features were: age, gender, time to ROSC, non-shockable cardiac rhythm, neurological exam on admission (presence of pupillary or corneal reflex and motor Glasgow Coma Score less than three), presence of an unfavorable CTH on admission, presence of malignant EEG patterns on EEG within the first 72 h, By including clinical, qualitative EEG review, and radiology data in the prediction model, we provide multimodal features to our outcome prediction model. The EEG reactivity expert scores from visual review were used as baseline and were considered "present" if there was an EEG response to either sound or pain stimuli.

First, we utilized a random forest classification model. We created 1000 rounds of bootstrap samples with replacement for training and testing sets, ensuring that subjects in the testing set were not included in the training set and vice-versa. Classification model training was performed using 5-fold cross-validation with 80% of subjects included in training and 20% in testing sets, i.e. the k-fold for random forest used five rounds of cross-validation for training and testing. In summary, the QEEG model was tested 5000 times using 1000 bootstrap rounds in five cross-validation folds. The training set was balanced using a randomized class equalization strategy to ensure that each fold had an equal number of subjects with good and poor outcomes (Fig. 1B, step "balance outcomes"). For example, if eighteen out of 40 subjects supposed to be included in a bootstrap round had good outcome, only 18 subjects with poor outcome would be included for training in that round, which would have a total of 36 subjects. This was done to ensure that the model voting strategy in the random forest would not be biased to always vote for the more frequent outcome instead of using the most relevant QEEG features. The MATLAB function TreeBagger was utilized to implement bootstrap-aggregated, i.e. bagged, decision trees. This method allows resam-

pling with replacement of the training set on each round of the bootstrap. The number of decision trees was fixed to 1000. We fixed the number of selected features to the 10 features with highest outcome prediction weights for each bootstrap round in the training set. These features would be included in the model used for testing. Second, we utilized a general linear model, i.e. penalized multinomial logistic regression with parameter optimization as a baseline prediction model (Tibshirani, 1996). This type of regression model utilizes variable selection and regularization in order to select a minimal set of features needed to optimize prediction accuracy. The penalized multinomial logistic regression used the same procedure described for the random forest model in regard to number of bootstrap rounds and training-testing set division. We report the mean AUC and Bayesian 95% credible interval (CI) for stimulus specific (sound and pain) and stimulus agnostic analyses (combined sound and pain) using 100 rounds of bootstrapping for both random forest and penalized logistic regression models. The visual EEG reactivity expert assessment was used as a baseline comparison method to the logistic regression and random forest model. We evaluated the performance of a visual EEG reactivity scored as "present" as a predictor of good outcome. The model sensitivity at a specificity threshold of 100% for good outcome prediction is reported for both random forest and penalized logistic regression models, i.e. no false-positive prediction.

Statistical calibration assesses how closely predicted probabilities of good outcome by a given model agree with the observed, i.e. actual, proportion of good outcomes (Osborne, 1991). The mean squared error of the prediction, i.e. Brier's score, indicates the difference from the predicted to observed outcomes. In this analysis, we report the mean Brier's score across the 1000 bootstrap iterations.

The random forest model outcome prediction output provides a probability of outcome at the subject level, and therefore can be used as a neurological recovery risk score. The probability is based on the proportion of trees in the ensemble that vote for a certain class. We report the cross-validated random forest classifier probability output scores obtained without resampling from the bootstrapping procedure. Difference in QEEG reactivity scores among good and bad outcome groups was evaluated using a t-student test. We also report the sensitivity at 100% specificity of both prediction methods to determine how frequently we can identify subjects with good outcome without having false-positive predictions. All statistical analyses were performed using MATLAB 2017a.

3. Results

Clinical and EEG data were available for 75 subjects. Twenty-five subjects were excluded through visual review due to presence of epileptiform discharges, burst-suppression, or major artifacts on EEG recordings at the time of the EEG reactivity testing battery. Fifty subjects were included in the final analysis, and 22 (44%) had arousal recovery prior to discharge, with 16 (76%) of them having a good outcome at six months (Table 2). Six subjects with poor outcome had arousal recovery and five were able to follow commands prior to discharge. A total of one-hundred and eighty-six stimulations were performed during EEG reactivity testing for all 50 subjects – 50 (26.8%) sound and 136 (73.2%) pain. Twenty-one (42%) subjects had EEG reactivity testing documented in more than one day (only information from first assessment analyzed). Electroencephalogram reactivity was considered present in 18 (36%) subjects (10 with good outcome and 8 with poor outcome) based on expert visual EEG review (Table 2). Malignant EEG patterns were present within the first 72 h of EEG monitoring in 37.5% of subjects with good outcome and 41.2% of subjects with poor outcome.

Table 2
Subjects characteristics.

	CPC 1–2 (N = 16)	CPC 3–5 (N = 34)	p
Age (years, mean, std)	50 (16)	58 (19)	<0.001
Female gender	37.5%	32.4%	NS
ROSC in minutes, (mean, std)	22 (23)	28 (21)	<0.001
Shockable Rhythm (VFib/VT)	56.3%	29.4%	NS
OHCA	87.5%	82.4%	NS
Motor GCS < 3*	88.3%	86.6%	NS
Pupillary reflex bilaterally absent*	40%	85.3%	0.001
Corneal reflex bilaterally absent*	71.4%	93.3%	0.04
Malignant EEG**	37.5%	41.2%	NS
EEG Reactive (visual expert review)	62.5%	23.5%	<0.01
CTH unfavorable*	11.1%	51.9%	0.03
WLST Poor Neurological Prognosis	0%	64.7%	<0.001
Arousal Recovery by discharge	100%	17.7%	<0.001
Following Commands by discharge	100%	14.7%	<0.001

ROSC: return of spontaneous circulation; VFib: ventricular fibrillation; VT: ventricular tachycardia; OHCA: Out-of-hospital cardiac arrest; GCS: Glasgow Coma Score; EEG: electroencephalogram; CTH: computerized tomography of the head; WLST: withdrawal of life-sustaining therapies; NS: non-significant.
* On admission.
** Malignant EEG: presence of burst-suppression, seizure, status epilepticus, myoclonic status epilepticus, or periodic discharges.

3.1. Outcome prediction using QEEG reactivity

In our initial analysis, we included as part of the random forest model several clinical, imaging, and EEG information by expert visual review other than QEEG reactivity features, i.e. admission clinical information (age, gender, OHCA, shockable rhythm, motor Glasgow Coma Score less than three, presence of pupillary or corneal reflexes), unfavorable admission head computerized tomography results, and presence of malignant EEG patterns. These features were not selected as the most predictive of outcome in the random forest model, therefore we only report results from models that exclusively included QEEG reactivity features.

The univariate AUC for outcome prediction using individual QEEG features for combined sound and pain stimuli is summarized in Fig. S1. Outcome prediction performance was equivalent between the random forest classification model compared to the baseline penalized logistic regression and expert review model for the stimulus agnostic analysis (mAUC = 0.8 vs. mAUC = 0.69 vs, mAUC = 0.69; differences were not statistically significant) (Fig. 2 and Supplementary Fig. S2). Sensitivity at the 100% specificity threshold for good outcome prediction in the stimulus agnostic analysis was 23% for random forest model and 8% for penalized logistic regression. The random forest model's calibration was equivalent to the penalized logistic regression model for stimulus agnostic and stimulus specific analysis for sound and pain stimuli (Supplementary Fig. S2; differences were not statistically significant).

3.2. QEEG reactivity features selected by the random forest model

The QEEG features selected for testing for the random forest model and penalized logistic regression models are summarized in Fig. S3. Quantitative EEG reactivity in the delta, theta, alpha, and spindle bands, i.e. change in power at these frequency bands, and entropy measures (non-linear energy operator, spectral entropy, and state entropy) were the most frequently included features for both stimulus specific and stimulus agnostic analysis.

3.3. QEEG reactivity as a neurological recovery probability risk score

We used a probabilistic approach to explore the performance of our random forest model in predicting good long-term outcomes at the subject level for each stimulus category separately (Fig. 3). Subjects in the good outcome group had higher neurological recovery probability scores for stimulus agnostic and the pain stimulus specific categories (p < 0.01 and p = 0.01, respectively). Neurological recovery probability scores generated using sound stimuli exclusively were not different between groups (p = 0.06). No speci-
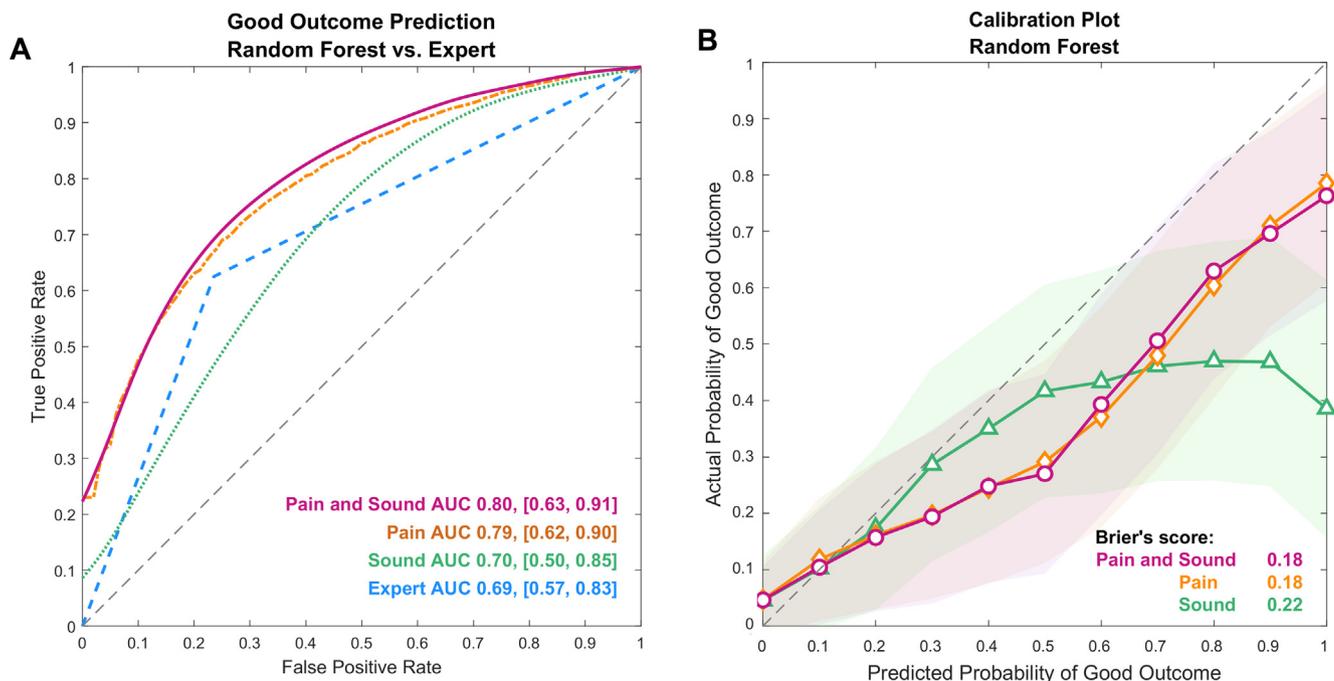

Fig. 2. A: Random forest model and expert visual review performance for good outcome prediction for different type of stimuli; B: Calibration plot of predicted vs. observed good outcome using a random forest model (mean score and standard deviation).
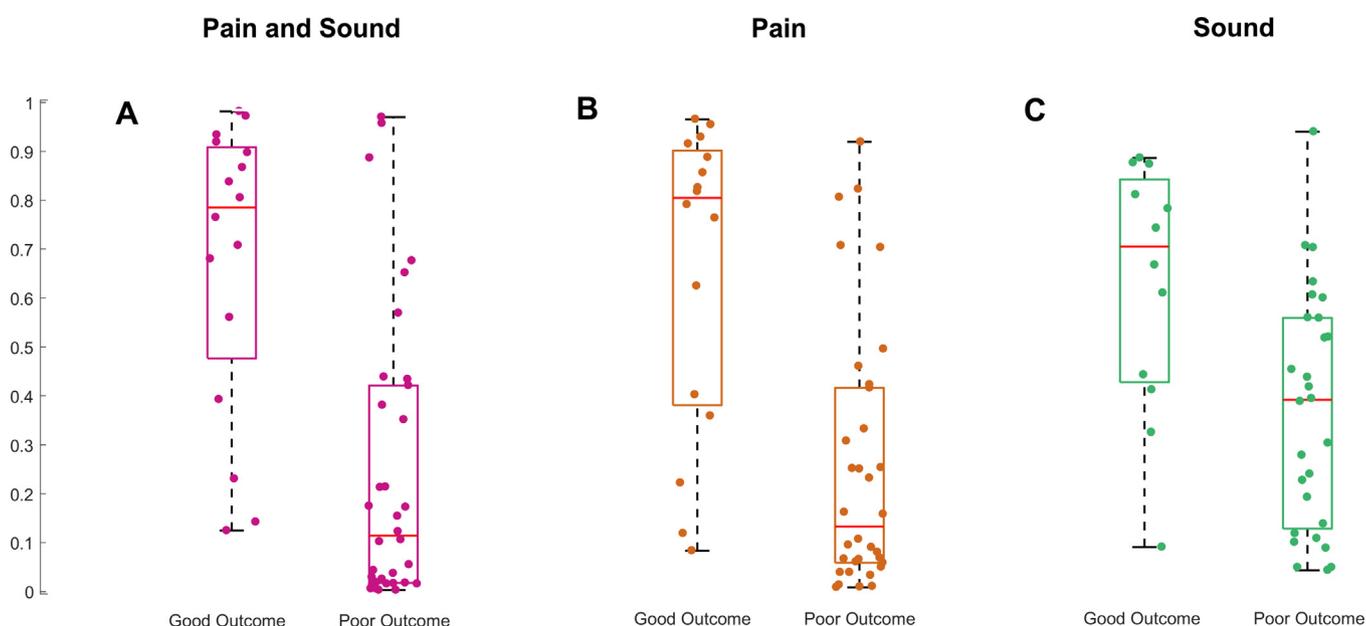
Pain and Sound   **Pain**   **Sound**

Fig. 3. Reactivity probability distribution for subjects with good and bad outcome for pain and sound (A, *p* < 0.01), pain (B, *p* = 0.01), and sound (C; *p* = 0.06) stimuli.

fic cutoff value could separate subjects with good and poor outcome with 100% specificity in any of the stimulus category groups. Fig. 4 illustrates the probability of good outcome obtained using quantitative EEG reactivity as a neurological recovery risk score along with the corresponding raw EEG signal and spectrogram during stimulation for four exemplary scenarios. In Fig. 4A, we illustrate the case of a subject with a low neurological recovery score to pain stimuli and bad outcome and in Fig. 4B a high neurological recovery score to pain and good outcome. The EEG reactivity response might be stimulus specific for some subjects. In Fig. 4C and D we show an example of a subject with good outcome who had diffuse attenuation of the background for three seconds after sound stimulation (Fig. 4C; neurological recovery score 0.7) and a much less prominent EEG response after pain stimulation (Fig. 4D; neurological recovery score 0.4).

4. Discussion

In this study, we demonstrate that machine learning models can analyze QEEG reactivity data to predict long-term neurological outcome in subjects with hypoxic-ischemic brain injury. Both the random forest model and the penalized logistic regression methods performed comparably to EEG reactivity assessments performed by an expert electroencephalographer. In addition, this data-driven approach can provide individualized estimates of good outcome using a neurological recovery risk score, allowing not just for the identification of a binary "change vs. no change" in the EEG signal to stimulation, but additionally providing the probability of neurological recovery as a continuous score. The neurological recovery score may be monitored over time during repeated reactivity assessments in parallel with the neurological exam.

We found that a combination of QEEG features assessing both spectral and entropy changes achieved best outcome prediction performance. Electroencephalogram reactivity by expert visual review is determined primarily by changes in frequency or amplitude. Quantitative features that discriminate changes in power spectra have also been used by other groups for QEEG reactivity determination or for outcome prediction – for both cardiac arrest and other types of neurological injury (Duez et al., 2018; Hermans et al., 2016; Johnsen et al., 2017; Liu et al., 2016;

Noirhomme et al., 2014). Entropy changes are not as easily translatable to standard EEG reactivity identified by expert visual review, and objective assessment of changes in signal complexity is not routinely pursued during standard EEG interpretation by experts. Assessment of entropy changes have only been utilized in QEEG reactivity testing on a previous study from our group using a mixed critical care population that included cardiac arrest subjects (Hermans et al., 2016). In that study, assessment of changes of relative entropy did not improve the EEG reactivity determination performance. In the present study, we focus on outcome prediction and not on prediction of agreement between expert EEG readers assessing EEG reactivity. We used an expanded QEEG feature set and pursued analysis using several types of entropy measures beyond relative entropy and utilized a more robust prediction model (random forest model). The random forest method has the advantage of avoiding overfitting despite inputting several variables and it also permits various approaches to feature selection. This semi-automated quantitative method can generate unbiased estimates of QEEG reactivity probability at both individual and group level, which could be used as a neurological recovery risk score.

Our methodological approach has several strengths. The use of machine-learning for outcome prediction using QEEG reactivity is novel and prevents known limitations in prognostication studies related to overfitting and correlated features. The algorithm design allowed for more interpretable results with individual-level outcome predictions despite a high dimensional dataset. The development of an algorithm that can evaluate stimulus-specific EEG responses is innovative. In addition, our statistical model was rigorous, and utilized five-fold cross-validation for training with an inner cross-validation split and 1000 round bootstrapping, generating more consistent results. The training sets were balanced according to neurological outcomes to avoid biasing the model to favor poor outcome predictions just because this was the most frequent outcome. The algorithm design in association with the heterogeneity of a data set originated from two different hospitals make our method more robust.

We were able to demonstrate that QEEG reactivity outcome prediction performance varied between QEEG feature classes and activation procedure categories. Electroencephalogram reactivity response is thought to represent the synchronization or desyn-
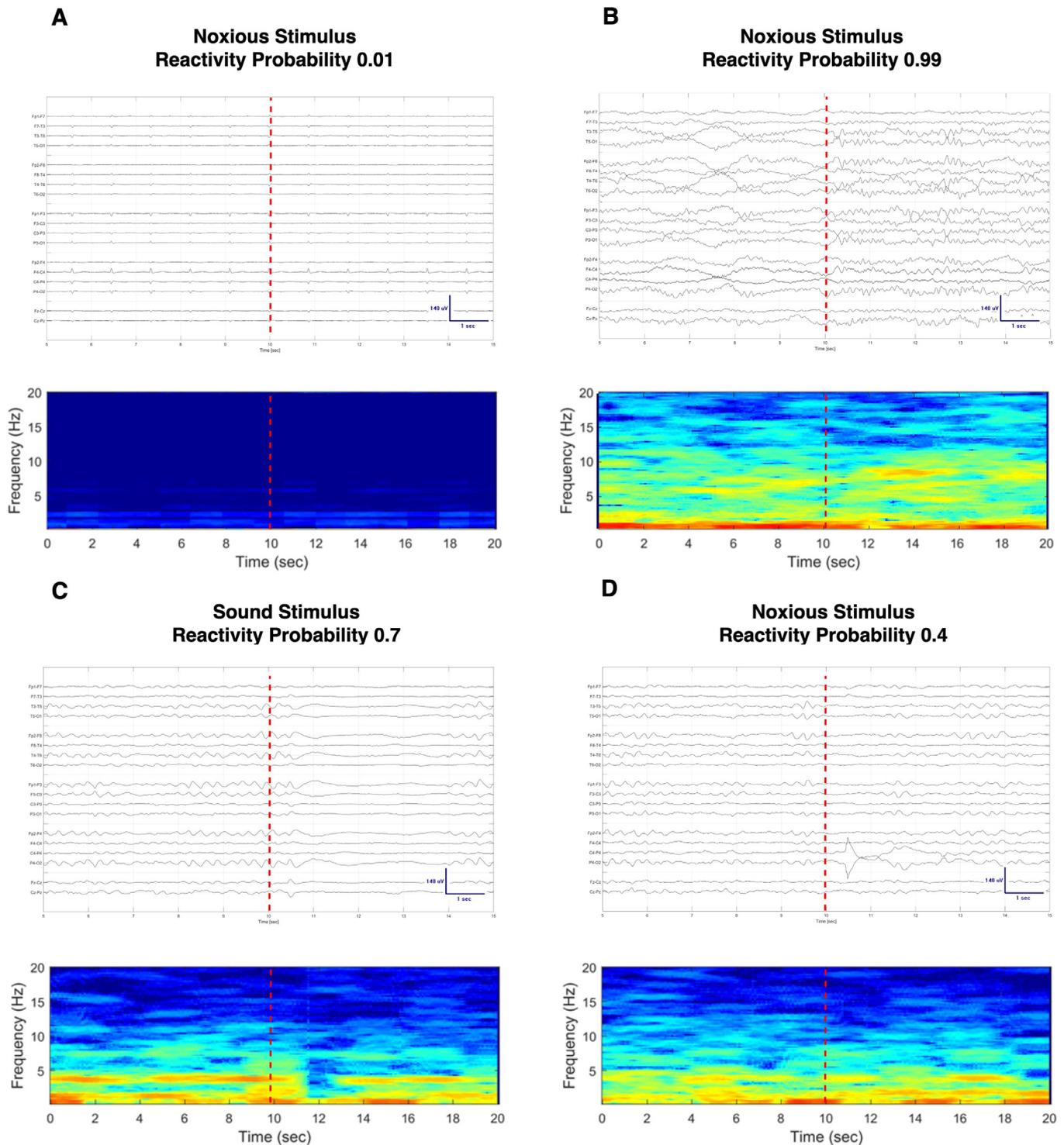
Fig. 4. Compressed spectral array display and raw EEG signal with corresponding probability of good outcome using QEEG reactivity as a risk score. A: subject had poor outcome and a low neurological recovery score; B: subject with good outcome and high neurological recovery score; C and D: subject with good outcome who had a higher neurological recovery score for sound vs. pain stimuli. There is a diffuse decrease in amplitude for three seconds after sound stimulation (C). No definite change on EEG is observable after noxious stimulation by visual inspection of the EEG recording (D).

chronization of cortical neuronal ensembles mediated by thalamo-cortical feedback loops (Hirsch et al., 2013; Pfurtscheller and Lopes da Silva, 1999). Therefore, it is not surprising that injury to different networks might generate distinct QEEG signatures. The QEEG reactivity features with best predictive performance in the univariate, random forest, and logistic regression model involved spectral power (delta, theta, alpha, and spindle bands) and entropy features (non-linear energy operator, spectral entropy, state entropy, per-

mutation entropy, and spectral entropy) (Figs. S1 and S3). In addition, while prediction performance for sound or pain stimuli as a group was equivalent (AUC 0.8 vs 0.69), at the individual level, we observed that some subjects might only have a more discernible EEG reactivity response, quantitative or by expert visual review, to one type of stimuli (Fig. 4). These differences could be related to dysfunction of specific neural circuits as sound and pain networks may have distinct patterns of cortical activation captured

on scalp EEG or different degrees of resilience to hypoxic-injury. In this context, different types of EEG reactivity signatures might be markers of arousal system integrity, which might be sufficient to allow command following or eye tracking, i.e. arousal recovery, but not translatable into good outcome as measured by the CPC scale.

EEG reactivity determination is often derived from a single assessment a day and its scoring frequently does not follow a systematic stimulation protocol using a structured clinical exam with sound and graded noxious stimulation (Amorim et al., 2018). In our study, predictions had good performance despite the fact that those were made with information from only the first EEG reactivity assessment. This finding reinforces previous literature indicating that EEG reactivity can be observed early and is a robust indicator of neurological recovery (Juan et al., 2015). However, patients who recover EEG reactivity during normothermia despite an unreactive EEG on hypothermia tend to have a favorable outcome (Juan et al., 2015). Therefore, performing serial EEG reactivity assessments when possible is recommended to increase the likelihood of identifying patients with potential for good outcomes.

Subjects with more severe brain injury often have malignant EEG and other clinical or radiological findings associated with poor outcome, including lack of EEG reactivity. In this study, we only included subjects without burst suppression and epileptiform discharges during the time of EEG reactivity assessment as fluctuation on bursts or epileptiform discharges frequency within the short 20-s windows used for QEEG reactivity analysis could lead to EEGs being considered reactive by chance. In clinical practice, there is no consensus about change in burst or epileptiform discharge patterns to stimulation being a predictor of poor outcome after cardiac arrest (Alvarez et al., 2013; Braksick et al., 2016). Additionally, several subjects with good outcome in our study had unfavorable clinical exam, qualitative EEG, and CTH findings. This might have contributed to the random forest model selection of QEEG reactivity features instead of clinical exam, qualitative EEG, and radiological features. This model only includes features that are most discriminative between groups with poor and good outcome. This finding highlights the importance of EEG reactivity assessment during multimodal prognostication in a population that might have contradicting outcome predictors.

This study has several limitations and future work is needed to advance QEEG reactivity assessments. First, the lack of statistical significance between the three methods evaluated (random forest model, penalized logistic regression, expert visual EEG review) might be due to a small sample size. Second, expert assessments were not blinded to other clinical data and prognostication is prone to self-fulfilling prophecies of withdrawal of life-sustaining therapies. Assessments employing multiple blinded raters' consensus could reduce bias from other prognostication sources that might have been available to the electroencephalographer. Third, our prediction model included clinical, imaging and EEG data routinely used in prognostication, however those features were not selected among the most predictive of long-term outcome. This could also have been due to the small sample size of the present study, which limits the number of factors the machine learning models can reliably learn to use. Fourth, we only evaluated EEG data obtained during the first EEG reactivity assessment as not all subject underwent daily serial assessments. EEG reactivity may re-emerge over time and we might have missed relevant information for outcome prediction that might have been biased by sedation and neuromuscular blockade use, targeted temperature management, and delayed recovery or worsening (Juan et al., 2015). Fifth, the EEG data was averaged across all scalp electrodes and different stimulations within an activation procedure category to reduce data dimensionality and allow a single EEG reactivity measure per QEEG feature. This loss of spatial and stimulus type information might have

decreased performance as reactivity responses may be more pronounced in specific brain regions or change depending on the type of activation procedure used (Fantaneanu et al., 2016; Hermans et al., 2016; Tsetsou et al., 2015). Sixth, the performance of this method needs to be evaluated in subjects who had malignant EEG patterns such as burst-suppression, periodic epileptiform discharges, and seizures as determination of EEG reactivity in these cases might have major repercussions in care and decisions about withdrawal of life-sustaining therapies. Future studies in cardiac arrest prognostication should pursue integration of all data relevant to prognostication in a synchronized way in order to allow the identification of confounders (e.g. sedation, neuromuscular blockade) and EEG and imaging signatures associated with recovery in neurological function at the bedside.

5. Conclusion

A machine-learning method using QEEG reactivity data can predict long-term neurological outcome in hypoxic-ischemic brain injury early. Analysis of individual cases suggests that these EEG reactivity signatures can be stimulus-specific for some subjects, reinforcing the need of utilizing different activation procedures and serial exams. Early and accurate assessment of neurological recovery biomarkers such as EEG reactivity are important steps to avoid misperceptions of prognostic expectations that can influence decision-making about withdrawal of life-sustaining therapies and potentially guide patient selection to interventional trials.

Declaration of Competing Interest

Nothing to report.

Author contributions

E.A., S.N., J.W.L., S.S.C., M.B.W. contributed to conception and design of the study. E.A., MvdS., S.N., M.M.G., and M.B.W. contributed to analysis of data. E.A., M.B.W. contributed to preparing the figures. E.A., MvdS., S.N., J.J., M.M.G., J.W.L., M.B.W. contributed to data acquisition. E.A., MvdS., S.N., M.M.G., J.J., J.W.L., U.O., S.S.C., M.B.W. contributed to drafting the text.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.clinph.2019.07.014.

References

Alvarez V, Oddo M, Rossetti AO. Stimulus-induced rhythmic, periodic or ictal discharges (SIRPIDs) in comatose survivors of cardiac arrest: characteristics and prognostic value. Clin Neurophysiol 2013;124(1):204–8.

Amorim E, Ghassemi MG, Lee JW, Westover MB. Dynamic quantitative EEG signatures predict outcome in cardiac arrest. Neurocrit Care 2016a;25 (Supplement 1):S158.

Amorim E, Gilmore EJ, Abend NS, Hahn CD, Gaspard N, Herman ST, et al. EEG reactivity evaluation practices for adult and pediatric hypoxic-ischemic coma prognostication in North America. J Clin Neurophys 2018;35(6):510–4.

Amorim E, Rittenberger JC, Baldwin ME, Callaway CW, Popescu A, Post Cardiac Arrest S. Malignant EEG patterns in cardiac arrest patients treated with targeted temperature management who survive to hospital discharge. Resuscitation 2015;90:127–32.

Amorim E, Rittenberger JC, Zheng JJ, Westover MB, Baldwin ME, Callaway CW, et al. Continuous EEG monitoring enhances multimodal outcome prediction in hypoxic-ischemic brain injury. Resuscitation 2016b;109:121–6.

Bokil H, Andrews P, Kulkarni JE, Mehta S, Mitra PP. Chronux: a platform for analyzing neural signals. J Neurosci Methods 2010;192(1):146–51.

Braksick SA, Burkholder DB, Tsetsou S, Martineau L, Mandrekar J, Rossetti AO, et al. Associated factors and prognostic implications of stimulus-induced rhythmic, periodic, or ictal discharges. JAMA Neurol 2016;73(5):585–90.

Bricolo A, Turazzi S, Faccioli F, Odorizzi F, Sciaretta G, Erculiani P. Clinical application of compressed spectral array in long-term EEG monitoring of comatose patients. Electroencephalogr Clin Neurophysiol 1978;45(2):211–25.

Duez CHV, Ebbesen MQ, Benedek K, Fabricius M, Atkins MD, Beniczky S, et al. Large inter-rater variability on EEG-reactivity is improved by a novel quantitative method. Clin Neurophysiol 2018;129(4):724–30.

Fantaneanu TA, Tolchin B, Alvarez V, Friolet R, Avery K, Scirica BM, et al. Effect of stimulus type and temperature on EEG reactivity in cardiac arrest. Clin Neurophysiol 2016;127(11):3412–7.

Giacino JT, Ashwal S, Childs N, Cranford R, Jennett B, Katz DI, et al. The minimally conscious state: definition and diagnostic criteria. Neurology 2002;58 (3):349–53.

Hermans MC, Westover MB, van Putten MJ, Hirsch LJ, Gaspard N. Quantification of EEG reactivity in comatose patients. Clin Neurophysiol 2016;127(1):571–80.

Hirsch LJ, LaRoche SM, Gaspard N, Gerard E, Svoronos A, Herman ST, et al. American clinical neurophysiology society's standardized critical care EEG terminology: 2012 version. J Clin Neurophysiol 2013;30(1):1–27.

Johnsen B, Nohr KB, Duez CHV, Ebbesen MQ. The nature of EEG reactivity to light, sound, and pain stimulation in neurosurgical comatose patients evaluated by a quantitative method. Clin EEG Neurosci 2017;48(6):428–37.

Juan E, Novy J, Suys T, Oddo M, Rossetti AO. Clinical evolution after a non-reactive hypothermic EEG following cardiac arrest. Neurocrit Care 2015;22(3):403–8.

Liu G, Su Y, Jiang M, Chen W, Zhang Y, Zhang Y, et al. Electroencephalography reactivity for prognostication of post-anoxic coma after cardiopulmonary resuscitation: a comparison of quantitative analysis and visual analysis. Neurosci Lett 2016;626:74–8.

Nielsen N, Wetterslev J, Cronberg T, Erlinge D, Gasche Y, Hassager C, et al. Targeted temperature management at 33 degrees C versus 36 degrees C after cardiac arrest. N Engl J Med 2013;369(23):2197–206.

Noirhomme Q, Lehembre R, Lugo Zdel R, Lesenfants D, Luxen A, Laureys S, et al. Automated analysis of background EEG and reactivity during therapeutic hypothermia in comatose patients after cardiac arrest. Clin EEG Neurosci 2014;45(1):6–13.

Osborne C. Statistical calibration: a review. Int Stat Rev 1991;59(3):309–36.

Perman SM, Kirkpatrick JN, Reitsma AM, Gaieski DF, Lau B, Smith TM, et al. Timing of neuroprognostication in postcardiac arrest therapeutic hypothermia. Crit Care Med 2012;40(3):719–24.

Pfurtscheller G, Lopes da Silva FH. Event-related EEG/MEG synchronization and desynchronization: basic principles. Clin Neurophysiol 1999;110(11):1842–57.

Rossetti AO, Oddo M, Logroscino G, Kaplan PW. Prognostication after cardiac arrest and hypothermia: a prospective study. Ann Neurol 2010;67(3):301–7.

Rossetti AO, Tovar Quiroga DF, Juan E, Novy J, White RD, Ben-Hamouda N, et al. Electroencephalography predicts poor and good outcomes after cardiac arrest: a two-center study. Crit Care Med 2017;45(7):e674–82.

Safar P. Resuscitation after brain ischemia. In: Grenvik A, Safar P, editors. Brain Failure and Resuscitation. New York: Churchill Livingstone; 1981.

Thomson DJ. Spectrum estimation and harmonic analysis. Proc IEEE 1982;70 (9):1055–96.

Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc 1996;58 (1):267–88.

Tsetsou S, Novy J, Oddo M, Rossetti AO. EEG reactivity to pain in comatose patients: importance of the stimulus type. Resuscitation 2015;97:34–7.

Westhall E, Rossetti AO, van Rootselaar AF, Wesenberg Kjaer T, Horn J, Ullen S, et al. Standardized EEG interpretation accurately predicts prognosis after cardiac arrest. Neurology 2016;86(16):1482–90.